Corpus-based Typological Research in Discourse and Grammar: GRAID and Multi-CAST

SCHNELL, Stefan
Centre of Excellence for the Dynamics of Language / University of Melbourne

SCHIBORR, Nils Norman
University of Bamberg

This volume gathers outlines of specialized corpus annotations for research into grammar and discourse from three different languages of Southeast Asia. The annotations implemented here follow the GRAID ('Grammatical Relations and Animacy in Discourse'; Haig and Schnell 2014) conventions. The papers document the implementation of these conventions in individual languages, which are to be included in the publication of the respective corpora in Haig and Schnell's (2015) Multi-CAST collection ('Multilingual Corpus of Annotated Spoken Texts', archived at the Language Archive Cologne). The three corpora have been annotated by the authors jointly with Stefan Schnell over the last year as part of a collaboration between the ARC Centre of Excellence for the Dynamics of Language and the ILCAA-based project *A collaborative network for usage-based research on lesser-studied languages*. In this introduction we provide an outline of the principles of GRAID annotations and their purpose in research on discourse and grammar.

Keywords: discourse and grammar, corpus annotation, corpus-based typology, referential choice, evolution of grammar

1. Introduction

This volume contributes to a project in corpus-based typological research in discourse and grammar that the authors have been developing in collaboration with Geoffrey Haig over the last eight years or so. The major concern of this project is to tackle standing problems in this area of research – for instance accounts of referential

SCHNELL, Stefan and SCHIBORR, Nils N. 2018. "Corpus-based typological research in discourse and grammar: GRAID and Multi-CAST". *Asian and African Languages and Linguistics* 12. pp.1–16. [Permanent URL: http://hdl.handle.net/10108/91145]

choice or the grammaticalization of agreement – through a corpus-linguistic approach to morphosyntactic cross-linguistic diversity. To these ends we have been developing a corpus annotation system (called GRAID, 'Grammatical Relations and Animacy in Discourse'; Haig and Schnell 2014) that captures key information on the form, function, and semantic properties of verbal arguments observable in running discourse.

The annotations have been designed with typological comparability in mind, and are applicable to natural languages of any morphosyntactic type. Yet given the huge diversity of human languages, annotators need to relate their language-specific analyses to respective typological interpretations when implementing our guidelines. Also, some fine-tuning with regards to individual language specificities is occasionally necessary. Hence, every single language corpus is accompanied by a documentation of the respective implementation of the GRAID guidelines in that language. The present volume gathers three such documentations from corpora of Southeast Asian languages. GRAID-annotated corpora are part of a multilingual corpus project (Multi-CAST, 'Multilingual Corpora of Annotated Spoken Texts'; Haig and Schnell 2015) hosted at the Language Archive Cologne (LAC), and the three language corpora represented in this volume are intended to be integrated into Multi-CAST in the near future.

The purpose of this introduction is to first provide an overview of research in 'discourse and grammar' which has emerged from the seminal work by Wallace Chafe and Talmy Givón in the 1970s, and the motivation for developing GRAID (Section 2). We then outline the principles of the GRAID scheme (Section 3), and how annotations can be quantified with regard to research questions in discourse and grammar (Section 4). We conclude this introduction with an overview of Multi-CAST, including the language corpora treated in this volume (Section 5).

2. Corpus annotations for discourse and grammar

Language is primarily used to talk about situations and the entities involved in them, and all comprehensive models of language deal centrally with the expression by linguistic structures of different types of situations. Of major concern here is the syntax of argument structure and the mapping of situation participants into syntactic positions (Fillmore 1977; Chomsky 1981; Dowty 1991; Goldberg 1995; Van Valin and LaPolla 1997; Bresnan et al. 2015; among many others). The considerable body of work in typological research has added to these general mechanisms of argument mapping the consideration of semantic saliency features such as animacy and definiteness (for instance Evans' 1997 famous distinction between 'role' and 'cast').

Since the 1970s, the tradition of discourse and grammar research pioneered by Wallace Chafe and Talmy Givón has been concerned with yet another aspect of the linguistic expression of situations, namely its integration into the linguistic context of discourse, that is, a text. A text has the characteristic of verbalizing interrelated situations, and it thus forms a coherent structure made up of individual

sentence constructions (Foley 2007). Of interest here is the formal expression of discourse referents, that is entities that participate in consecutive situations throughout a discourse. This choice of expression has been found to be determined by various factors pertaining to discourse structure and the activation state of referents (Chafe 1976; Prince 1981; Ariel 1988; Gundel et al. 1993), but also to be dependent on semantic and syntactic factors (Du Bois 1987; Du Bois et al. 2003, and contributions therein). Moreover, referential choice has in turn been claimed to have a crucial impact on the diachronic development of grammatical systems encoding situations and their participants (Givón 1976, 1979; Hopper 1998; Du Bois 1987; Ariel 2000). This view can be characterized as the functionalist, or usage-based, approach to argument structure, and grammar in general. In this view, grammatical systems emerge from language use in discourse processing, comprising production and comprehension.

The Chafe-Givón tradition has shown a huge range of corpus-based studies from typologically diverse languages, and has in this sense been crucial in paving the way for the contemporary developments in corpus-based typology that our research is drawing on (see for instance Bickel 2003; Noonan 2003; and Du Bois et al. 2003 for more recent pioneering studies in the wider field of discourse and grammar). While Bickel (2003) provides a detailed account of how they analysed and counted different types of referring expressions in their corpora (as does Kumagai 2006 in their study of Preferred Argument Structure, based on Chafe's 1980 Pear stories from English), most researchers working in this tradition have not made their procedures and/or corpus data available to scrutiny by other scholars. Overall, in spite of the numerous achievements in discourse and grammar research across languages, no efforts have been made thus far to develop a typological database (like WALS or GramBank for structural features) for this research agenda.

GRAID ('Grammatical Relations and Animacy in Discourse'; Haig and Schnell 2014) and related developments have been initiated not only with the aim of filling this gap, but also of improving on different methodological and conceptual aspects of the Chafe-Givón tradition. We follow this tradition in regarding the exploration of language use and discourse processing in the area of referentiality as a vital enterprise, necessary for coming to an understanding of the discourse basis of grammatical systems. But we take issue with some of its premises, which have often led to mere confirmations of previously established hypotheses and an abundance of post-hoc accounts of observed patterns in referential choice. Our goal is thus to establish a research program within the emergent field of corpus-based typology that allows for explicit and rigorous testing of hypotheses pertaining to referentiality in discourse and grammar. The discourse data it draws from should be as unbiased as possible. A more mundane goal in this area is simply the expansion of the available discourse data into a typologically larger and more diverse sample of languages, and also the differentiation to more text categories, beyond the canonical Pear stories which have traditionally formed the basis of studies

in discourse and grammar.[1]

3. GRAID annotations for investigations in discourse and grammar

GRAID has been developed for the purposes of research in the area of referentiality at the syntax-discourse interface, as pioneered by Wallace Chafe, Talmy Givón, and many others working in the tradition of discourse and grammar. The most central research question concerns the choice of referring expressions, that is whether syntactic arguments with different functions are realized as full noun phrases (NPs) or pronouns, or left unexpressed.

Hence, the core of GRAID annotations captures the form and function of syntactic arguments, and, given their widely observed significance, their person and animacy features. Other clause constituents, like predicates and adjuncts, are registered only coarsely. In addition to clause constituents, annotations also note clause boundaries and types of dependent clauses (relative, complement, and adverbial clauses), as well as illocutionary force, and whether a clause constitutes (part of) direct speech. The full conventions are web-published as Haig and Schnell (2014); our purpose here is merely to explain the basic ideas of the annotation system. The following example from Schiborr's (2015) corpus of autobiographical texts from English illustrates the basic GRAID glossing:

(1) ENGLISH

a.	I	went	along	with	this	old	man,	Mr. Brown,
##	pro.1:s	v:pred		adp			np.h:obl	
b.	he	was	a	nice	old	man.		
##	pro.h:s	cop				np.h:pred		
c.		Used	to	have	a	team of four great horses.		
##	0.h:a			v:pred		np:p		

For the determination of syntactic functions, GRAID conventions follow Andrews' (2007) conception of a semantic transitive prototype, and generalizing over the form of argument encoding. Thus in English, a clause headed by the verb *kill* or *smash* would describe a prototypical transitive event with prototype agent and patient, where the former is encoded as a pre-verbal NP triggering agreement on the verb in the 3^{rd} person in present tense, and the latter is encoded as a post-verbal NP. Any syntactic argument that shows the same encoding as a prototypical agent or patient in a transitive clause are said to bear A or P function, respectively. Any clause without either A or P argument

[1] It is worth noting here that we are not in principle opposed to using text data elicited by means of stimuli like the Pear Film. It is plainly obvious that for some research questions, for instance referential density, control of content and opportunities of verbalisation across corpora is absolutely vital. Yet, we argue that the focus on such data to the exclusion of other types of texts does bear significant caveats, as is clear from our critical assessment of preferred argument structure theory in Haig and Schnell (2016, see in particular the online appendix).

is deemed intransitive. It may contain a single core function S, or possibly additional oblique functions. S is also the function ascribed to the subject of a non-verbal clause.

Hence, the 1^{st} person pronoun in (1a) bears S function in an intransitive clause, since the second argument here bears oblique rather than P function. The oblique argument is encoded by a preposition, which is noted in the annotations. Example (1b) is a non-verbal clause with a nominal predicate, and the function of the 3^{rd} person pronominal subject is S. In (1c), the construction is transitive, in that both arguments of a construction with *have* show the same encoding as prototypical agent and patient, and thus the two respective argument functions here are A and P, regardless of the fact that the states of affairs expressed here is fairly atypical in terms of semantic transitivity.

The use of syntactic functions S, A, and P rather than of a language-specific category like 'subject' is required in order to facilitate cross-linguistic comparison of corpus annotations and related findings. An alternative of these categories, as defined by Andrews (2007), would be the use of generalized semantic macro-roles, like the ones proposed in role-and-reference grammar (Van Valin and LaPolla 1997), or in Bickel (2011). A major motivation for the use of Andrews' framework is the relative ease with which core-argument functions can be identified and delimited, and the fact that much of the research, for instance the proposals regarding preferred argument structure (Du Bois 1987), has been framed in terms of syntactic functions rather than semantic macro-roles. Modification of this conception is, however, required where languages show alternations between multiple transitive constructions with systematic mapping alternations of core arguments, for instance in so-called Philippine-type languages with symmetrical voice systems (Riesberg and Primus 2015; Foley 2008). Here, only a macro-role approach is applicable, as has been done in Brickell and Schnell (2017). This approach has not been implemented in any of the language corpora presented in this volume, so with considerations of space in mind we will not elaborate on this point here.

As for the form of arguments, the identification of full noun phrases as well as pronouns is quite straightforward in English. In languages that do not possess pronouns for 3^{rd} person reference and use demonstratives or some other pro-form instead (see Bhat 2004), those pro-forms that are used like definite pronouns (Lyons 1968) in a language like English are glossed ⟨pro⟩ in GRAID, but receive an additional tag to indicate that it is, for instance, a demonstrative (e.g. ⟨dem_pro⟩). Similar considerations apply to demonstratives used pronominally even in a language like English or Teop (see Mosel and Schnell 2015), and other "special" pronominal forms such as relative pronouns. This procedure enables global cross-linguistic comparison with regards to a comparative category of ⟨pro⟩, as outlined above, as well as permitting consideration of different form types within this broader category.

One notoriously contentious category in the GRAID system is that of zero: in (1c), we annotate a zero form that fulfils the function of an A argument in a transitive construction headed by have. The first point to clarify here is that 'zero' in GRAID –

and in all the documentations included in this volume – is equivalent to an unexpressed referent on clause level; it is as such fundamentally different from paradigmatic zeroes in morphological constructions. Furthermore in GRAID, zero arguments are quite restricted and generally subject to the following conditions:

1. the predicate (i.e. the verb in most cases) must license the argument in question.
2. The intended referent must be specific and retrievable from the discourse context. This essentially means that semantic entailment (Dowty 1982) is not a sufficient criterion to assume a zero argument (see also for instance Dalrymple 2001).
3. The predicate-argument construction in question must not systematically suppress the argument function in question, so that it can essentially never assume a form other than zero. Thus, for instance English to-complements would not be regarded as taking zero S or A arguments. The rationale behind this procedure is that in these structures, we do not find a contrast between zero and other forms of expression, so that in practice speakers do not have a choice of form in the way they have in contexts like (1c), where a pronoun (or, in principle at least, a full noun phrase) could have been used instead (see also Bickel 2003 for these considerations).

While we clearly intend to restrict this category to referential zeroes, the third criterion bears some complications: while the technical quantitative considerations regarding the alternation of forms is undisputable, the non-consideration of suppressed argument functions seems to give an odd impression with regards to more global ideas about the implicitness of discourse, in particular in languages that make intensive use of desententialized/nominalized clause constructions, like infinitive, participle, or nominalized clauses. Here, a value for referential density may turn out to be relatively high, merely due to the fact that implicit referents can never be realized overtly in the constructions attested in a given text. For this reason, in order to evaluate the global properties of more or less implicit discourse, it is required to consider the relative proportion of argument-suppressing constructions and respective suppressed argument functions. The same considerations apply to relative clause constructions with gapping as a regular relativization strategy, so that again no overt form can ever occur in this specific syntactic configuration.

We therefore plan to adopt a glossing practice that is in a sense similar to that of pro-forms: we gloss contrastively suppressed arguments as zero ⟨0⟩, and (optionally) assign a different symbol ⟨f0⟩ to implied participants in non-finite clauses. The latter signals the suppression as such, as well as the type of construction that the argument is implied in. For instance, a forced gap in a participial clause construction is glossed ⟨pt_f0⟩. This practice allows the evaluation of both the number of true contrastive zeroes as well as all implied participants. Note that in contrast to pronouns, we would generally apply the more restricted count to zeroes, following Bickel (2003). This practice has not been adopted yet in our corpora. Corpora where this will

be of particular relevance are the ones from the Tibeto-Burman languages Burmese (annotation in progress by Pavel Ozerov) and Jinghpaw (Kurabe, this volume) since these are relatively rich in non-finite constructions.

As for the semantic properties of arguments, the first distinction is that in person: we gloss whether an argument has a 1^{st} person ⟨.1⟩ or 2^{nd} person ⟨.2⟩ referent. These essentially mean reference to speaker or addressee. Where their respective forms, such as 2^{nd} person pronouns, are used for other kinds of reference, these are either excluded – as in the case of generic statements which essentially constitute meta-linguistic commentary – or need to be specially noted by the annotator. Note also that, where a 1^{st} person pronoun is used to refer to a narrator outside the narrated reality, it is categorically excluded from glossing, as are any clauses that represent meta-linguistic commentary rather than narrative clauses (Haig and Schnell 2014; Bickel 2003). Where a narrator is at the same time a participant in the narration, as in autobiographical narratives, 1^{st} person references are glossed. Likewise included are instances of direct speech with 1^{st} and 2^{nd} person reference to characters in the narrative (cf. Haig and Schnell 2014: 24).

Arguments in the 3^{rd} person are not overtly glossed for person, but we do indicate when the referent is human ⟨.h⟩; this can be seen in all three examples in (1) with NP, pronoun, and zero arguments, respectively. We also capture reference to anthropomorphized beings, for instance in mythical stories, which are then glossed ⟨.d⟩ (standing for 'deity', which is a common human-like being in narratives from the Middle East and Southeast Europe in our corpus) rather than ⟨.h⟩. Among anthropomorphized or human-like beings we count those entities that are capable of speech and thought. This allows us to count both humans proper as well as human-like beings in a unified category, or differentiate them, depending on what appears more appropriate in a given specific research context.

It should be noted that GRAID provides the possibility to leave linguistic material unconsidered. This would typically be required for structures that are either incomplete or clearly not well-formed (i.e. production mistakes), or for which an adequate analysis is impossible to come by at a given stage of investigation of the language in question. This is an important feature, in particular given that GRAID is intended to be used mainly on corpora from hitherto under-researched languages. As a rule of thumb, the amount of discourse left unannotated should not exceed 10% of all utterance units (Haig and Schnell 2014: 28).

A final remark concerning the design and implementation of glosses and the number of categories considered in GRAID: the reader will have noted that we use multi-barrelled expressions as glosses on a single level of annotation, essentially a tier in ELAN or a field in Toolbox. Our motivation for keeping these complex gloss words, rather than creating individual annotation tiers for each domain, derives mainly from practical considerations from the point of view of the annotator: GRAID annotations are typically undertaken on hitherto under-researched languages, which means they

are for the most part done manually, and it seems easier for most annotators to handle only a single tier. Furthermore, the GRAID gloss words are reminiscent of traditional morphemic glosses which makes them easy to conceptualize. It should be noted that the different domains of glossing (form, semantics, syntactic function) are clearly separated via unique delimiters, and in fact glosses can be easily dissected, for instance for subsequent work in a spreadsheet or in R. It is also important to note that, although we align GRAID glosses with grammatical words, they essentially pertain to phrases on clause level; see the GRAID manual (Haig and Schnell 2014) for details on how to handle constituents of complex phrases. Likewise, our aim to keep the number of categories as small as possible is motivated primarily by practical considerations pertaining to manual glossing. Also, our relatively coarse distinctions, for instance the (non-)human distinction in ontological classes, has so far proven to be sufficient for most purposes. Where researchers envisage that the relevant distinctions are clearly insufficient, they may introduce further categories, say for instance ⟨`.a`⟩ (in the semantic feature slot) for non-human animates.

Before we continue with an outline of the possibilities of quantifying GRAID glosses for specific research questions, it is worth mentioning a recent further development of our corpus annotations that facilitates research in referential choice. We have recently been developing an additional annotation system which we call RefIND ('Referent Indexing in Natural-language Discourse'; Schiborr, Schnell and Thiele 2018). The following is our text example (1) from above, this time with referent indexes added on an additional line under the GRAID annotations:

(2) English

a.	I	went	along	with	this	old	man,	Mr. Brown,
##	pro.1:s	v:pred		adp			np.h:obl	
	0000						0023	
							new	
b.	he	was	a	nice	old	man.		
##	pro.h:s	cop				np.h:pred		
	0023							
c.		Used	to	have	a	team of four great horses.		
##	0.h:a			v:pred		np:p		
	0023					0024		
						new		

The principles of referent index annotation are relatively simple, in that it merely requires the assignment of a unique numerical identifier to each referring expression that the annotator deems referential. The true challenge lies with the latter part, namely the decision as to whether a given expression is referential or not. We provide guidelines in the RefIND manual (Schiborr, Schnell and Thiele 2018) available from the Multi-CAST website.

We combine the RefIND annotations with a drastically simplified version of Riester

and Baumann's (2017) RefLex conventions. Specifically, we distinguish only between ⟨new⟩ and ⟨bridging⟩ references, with the former being roughly equivalent to Prince's (1981) 'brand-new' and 'unused', and the latter covering all instances of any 'evoked' referent, whatever the cause of its evocation. The purpose of these two layers, RefIND and the simplified RefLex, is essentially the same as that of RefLex itself, that is capturing information on information status comprising the introduction of different types of referents into the universe of discourse, as well as the tracking of referents through a discourse. The reason to not annotate referring expressions for specific information status is two-fold: for one thing, annotation by indices can be done much quicker and with greater ease once the basic decision about the existence of a particular discourse referent has been made. For another, this type of indexing facilitates investigations into issues of referent introduction and tracking without imposing specific kinds of information-related categories and units. For instance, we can measure the anaphoric distance between two mentions in a discourse in terms of number of clause units, words, morphemes, competing referents, or even time (at least roughly, depending on time-alignment of different levels of annotation); we can also incorporate different syntactic levels, for instance relationships between subordinate clauses and independent clauses. These additional levels of annotation have not yet been implemented in all Multi-CAST corpora, and not in the corpora presented in this volume, but are planned to be finalized in the near future.

4. Quantifying and analyzing basic GRAID annotations in discourse and grammar research

Once all arguments have been annotated as described above, the GRAID annotations (⟨pro.1:s⟩, ⟨pro.2:s⟩, . . . , ⟨np.h:a⟩, . . .) can be queried and quantified with regards to questions on discourse and grammar. We will here outline a number of relatively simple searches for the quantification of proportions.

A first very simple example concerns the relative humanness of core argument functions in transitive clauses: it has often been mentioned in the typological literature that typical transitive events involve a human actor acting upon a non-human or inanimate patient or theme (Comrie 1989). This should be reflected, at least roughly, in the respective proportion of human to non-human referents in the A and P arguments of transitive clauses. GRAID annotations can be used straightforwardly to determine these proportions by adding up all glosses with a function gloss ⟨:a⟩, then only those instances of ⟨:a⟩ with human reference, essentially all expressions with 1st person ⟨.1⟩, 2nd person ⟨.2⟩ and 3rd person human ⟨.h⟩ or human-like ⟨.d⟩ glosses, and then dividing the latter by the former.

We can also turn the proportions around if we were interested in the propensity of a human versus a non-human referent to occur in each argument role. If we restrict, for the sake of simplicity, this example to core argument functions S, A, and P, then

what we would do is first add up all gloss words with the function glosses ⟨:s⟩, ⟨:a⟩, and ⟨:p⟩ *plus* semantic glosses ⟨.1⟩, ⟨.2⟩, ⟨.h⟩, and ⟨.d⟩. To these totals we can then relate each individual sum of the relevant individual subcategories of all human S, all human A, and all human P, respectively. It is crucial to note that this perspective is fundamentally different from the first one, although it draws from the same quantities of data. It may be potentially relevant for questions concerning the processing of argument structure, so that where a hearer/reader comes across a NP with a human head noun, they may venture guesses as to the grammatical function of this NP in a clause. This is a question relevant for neurotypological research on the processing of ergativity, as in Bickel et al. (2015), though these authors do not find any effect of human versus non-human reference on the processing of unmarked (i.e. not case-marked) NPs in the split-ergative language Hindi.

A further example of a research question to which GRAID-annotated corpus data can potentially contribute is that of referential density (RD, Bickel 2003). Referential density is a conceptualization of the explicitness of speakers about discourse referents. We can determine the number of all arguments in a given text or corpus by searching for all syntactic function glosses (⟨:a⟩, ⟨:s⟩, ⟨:p⟩, ⟨:obl⟩, ⟨:other⟩, etc.) and then do the same for all instances with ⟨np⟩ or ⟨pro⟩ glosses (excluding ⟨0⟩), and then divide the former sum by the latter. This would yield the regular RD value, as investigated by Bickel (2003). By the same token, one can likewise determine the value of lexical RD, as investigated by Stoll and Bickel (2009); here, only the ⟨np⟩ glosses would be added up in the second step, and this sum then divided by the overall number of argument functions attested in a text or corpus. Corpus investigations into referential density seek to arrive at empirically grounded explanations for tendencies in individual languages to leave arguments unexpressed, a possibility captured under the label 'pro-drop' in the generative tradition (see Neeleman and Szendröi 2007). Obviously, the tendency to deploy a higher proportion of overt forms can be due to the frequent use of pronouns, which is implied in the term 'pro-drop'; it could, however, also be due to a frequent use of NPs, hence the differentiation in two types of RD values.

It is important to note at this point that although GRAID-annotated corpora can potentially contribute to RD research, they would have to fulfil further requirements, and the counts would have to be refined. Both Bickel (2003) and Stoll and Bickel (2009) use Pear Film retellings for their comparative studies on RD. This is a defining feature of their studies, since the use of overt expressions, and in particular full NPs, will crucially depend on the number of different entities that speakers mention in a given text, which is thus merely a question of content. Hence, in order to determine systematic differences in the behaviour of speakers across languages, the content needs to be kept relatively stable. A glance at the data overview in Bickel (2003) reveals that the content of Pear stories may possibly vary quite drastically, given the noticeably different lengths of texts in the database; similarly, Kumagai's (2006) counts reveal considerable variation in the number of human referents mentioned by the English

Pear story narrators in Chafe's (1980) corpus. Yet the important point here is that participants in this text elicitation experiment have the same number of *opportunities* to verbalize referents. In sum, any study on referential density should be undertaken based on Pear story text data, as comparisons of RD across corpora of uncontrolled text data are basically meaningless. Thus, while GRAID annotations can serve as the basis for determining RD values in a text, the Multi-CAST collection is not immediately amenable to this kind of research.

The question as to whether a noun phrase is used in favour of either pronominal or zero realisation of an argument has been quite central to the discourse and grammar research tradition. It is, for instance, central to Ariel's (1990) accessibility theory. To investigate the role of this choice in referent tracking, the extended annotation system combining GRAID and RefIND is required, and it presupposes a relatively elaborate mechanism of querying and analysis. Here, I will focus on one aspect of the distinction, namely the preference or dispreference of specific argument function to be realized by a noun phrase. Two claims have been prominent within the discourse and grammar tradition, namely Chafe's (1994) light subject constraint and Du Bois' (1987) theory of preferred argument structure (PAS).

The former predicts that both A and S are unlikely to be realized by full noun phrases. The latter makes the similar, but crucially different, prediction that only the realization of A by a full noun phrase is unlikely and in fact avoided, whereas both S and P are free to host full noun phrases. Both hypotheses underscore a crucial interrelation between discourse structure and sentence structure: while the mapping of participant roles and their formal realisation seem to each belong to either of the two different domains, quantitative investigations of discourse data suggest a subtle interrelation between the two, so that syntax would in fact not only link to semantic-conceptual event structure but also to discourse structure. Moreover, the two postulated patterns in discourse blatantly mirror patterns of grammatical structure, so that the light subject constraint corresponds to accusative and PAS to ergative alignment. As for the latter, Du Bois (1987) claims the existence of a diachronic relationship between the two, so that the freedom of S and P to host full noun phrases ultimately leads to ergative grammar. This is a crucial characteristic of the general emergentist view on grammar as represented in the discourse and grammar tradition.

To assess the claims associated with PAS, GRAID'ed corpora can be investigated by focusing investigations of lexical referential density on specific functions of interest. Hence, we can determine the level of full noun phrase expressions in A function by adding up all glosses ⟨`:a`⟩, then adding up only those ⟨`:a`⟩ glosses that also contain a form gloss ⟨`np`⟩, and finally dividing the latter by the former. The same can be done for the S and P functions, and then the proportions can be compared. In Haig and Schnell (2016), this has been done for five of the Multi-CAST corpora. The authors find more support for a light subject constraint than PAS, and crucially they also find a correlation between the likelihood of each role to host human referents and their dispreference

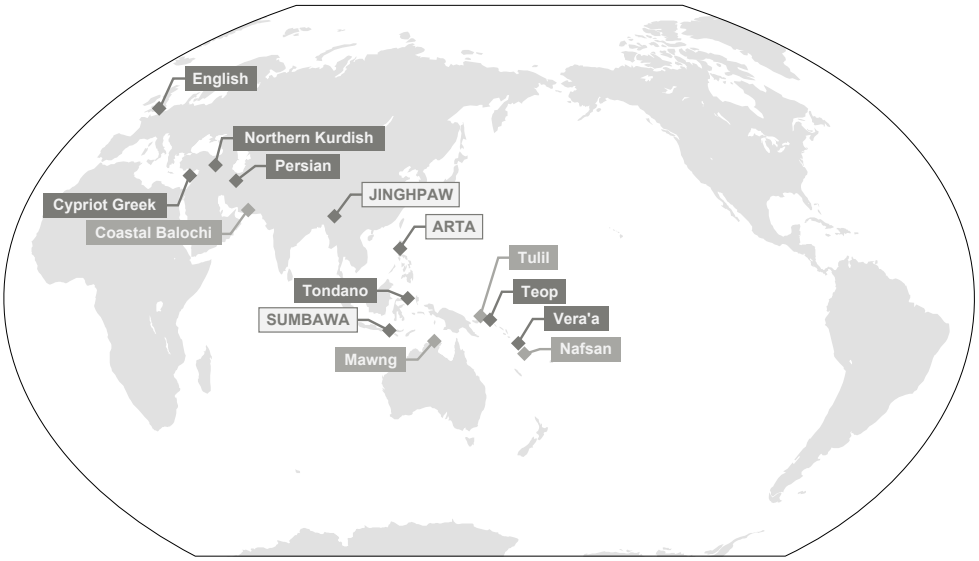

Fig. 1 Multi-CAST corpus locator map.

for full noun phrase realisation, so that the tendency for A (and in most corpora also S) to be realized by a pronoun or zero rather than a full noun phrase appears to be epiphenomenal of humanness, rather than to be a reflection of a discourse-structural linking of argument structure.

5. Expanding the empirical bases for cross-corpus typological research

As indicated above, the research questions exemplified here have a long and elaborate tradition in typologically oriented functional linguistics. Despite the great achievements of this research tradition, they have never resulted in a cross-linguistic database that could be used by other researchers to assess their claims and/or develop further usage-based models of grammar grounded in discourse. While another tradition in the area of variationist sociolinguistics does promote the explicit encoding of relevant information for statistical analysis, the data is often not accessible, and is typically separated from the actual corpus data, usually in the form of coding in spreadsheets. Relevant coding can of course still be related back to respective corpus instances (these are usually part of the coding in the spreadsheet), but the relationship needs to be re-established every time a given instance is scrutinized.

A central goal of the GRAID and Multi-CAST initiative is make a contribution towards closing this gap little by little. The fact that we capture relevant information in

the form of corpus annotations makes instances of argument realisation patterns much better retrievable from corpus data. Moreover, not only all our annotated corpus data, but also all quantitative data as well as the overall annotation guidelines and individual language-specific documentation of their implementation are made available as part of Multi-CAST, so that every step in any GRAID-based corpus study is transparent to other scholars.

A standing challenge to our efforts is the typological broadening of our corpus database. While the design of our annotation system does seem to facilitate relatively easy and efficient manual annotation, the finalization of each language-specific corpus is a major effort that takes up a substantial amount of resources to achieve. So far, the Multi-CAST collection features seven GRAID-annotated corpora. Four further corpora (from Mawng, Nafsan, Tulil, and Coastal Balochi) are in essence finished and will be published online in the near future.

This volume contains documentation of GRAID implementations in three further corpora from languages of Southeast Asia. These corpora have been annotated in collaboration with Schnell over the course of 2017, when researchers from the Tokyo University of Foreign Studies (TUFS) visited the Melbourne node of the Centre of Excellence for the Dynamics of Language (CoEDL). A corpus of traditional folk tales from the Tibeto-Burman language Jinghpaw, spoken in Myanmar, has been annotated by Keita Kurabe. Another corpus of narrative texts from Arta, an Austronesian Luzon language spoken in the Philippines, has been annotated by Yukinori Kimoto. A further Austronesian language corpus comes from Sumbawa, spoken on the island of the same name in Indonesia; it is annotated by Asako Shiohara.

References

Andrews, Avery. 2007. “The major functions of the noun phrase”. In Timothy Shopen (ed.) *Language typology and syntactic description* [Vol. 1]. 2nd ed. Cambridge: Cambridge University Press. pp.132–223.

Ariel, Mira. 1988. “Referring and accessibility”. *Journal of Linguistics* 24(1). pp.65–87.

———. 2000. “The development of person agreement markers: From pronouns to higher accessibility markers”. In Michael Barlow and Suzanne Kemmer (eds.) *Usage-based models of language*. Stanford: CSLI. pp.197–220.

———. 2014[1990]. *Accessing noun-phrase antecedents*. London: Routledge.

Bhat, D. N. S. 2004. *Pronouns*. Oxford: Oxford University Press.

Bickel, Balthasar. 2003. “Referential density in discourse and syntactic typology”. *Language* 79(4). pp.708–736.

———. 2011. “Grammatical relations typology”. In Jae J. Song (ed.) *The Oxford handbook of language typology*. Oxford: Oxford University Press. pp.399–444.

Bickel, Balthasar, Alena Witzlack-Makarevich, Kamal K. Choudhary, Matthias Schlesewsky and Ina Bornkessel-Schlesewsky. 2015. “The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking”. *PlosOne* 10(8). (DOI:10.1371/journal.pone.0132819).

Bresnan, Joan, Ash Asudeh, Ida Toivonen and Stephen Wechsler. 2015. *Lexical-functional syntax*. 2nd ed. Hoboken, NJ: Wiley-Blackwell.

Brickell, Timothy and Stefan Schnell. 2017. “Do grammatical relations reflect information status? Reassessing preferred argument structure theory against discourse data from Tondano”. *Linguistic Typology* 21(1). pp.177–208.

Chafe, Wallace. 1976. “Givenness, contrastiveness, definiteness, subjects, topics, and point of view”. In Charles N. Li (ed.) *Subject and topic*. New York: Academic Press. pp.25–55.

——— (ed.). 1980. *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex.

———. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: The University of Chicago Press.

Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Comrie, Bernard. 1989. *Language universals and linguistic typology*. 2nd ed. Oxford: Blackwell.

Dalrymple, Mary. 2001. *Lexical-functional grammar*. San Diego: Academic Press.

Dowty, David R. 1982. “Grammatical relations and Montague grammar”. In Pauline Jacobson and Geoffrey K. Pullum (eds.) *The nature of syntactic representation*. Dordrecht: Springer. pp.79–130.

———. 1991. “Thematic proto-roles and argument selection”. *Language* 67(3). pp.547–619.

Du Bois, John. 1987. “The discourse basis of ergativity”. *Language* 63(4). pp.805–855.

Du Bois, John, Lorraine Kumpf and William J. Ashby (eds.). 2003. *Preferred argument structure: Grammar as architecture for function*. Amsterdam: John Benjamins.

Evans, Nick. 1997. “Role or cast? Noun incorporation and complex predicates in Mayali”. In Alex Alsina, Joan Bresnan and Peter Sells (eds.) *Complex predicates*. Stanford: CSLI Publications. pp.397–430.

Fillmore, Charles J. 1977. “Scenes-and-frames semantics”. In Antonio Zampolli (ed.) *Linguistic structures processing*. Amsterdam: North-Holland. pp.55–81.

Foley, William A. 2007. “A typology of information packaging in the clause”. In Timothy Shopen (ed.) *Language typology and syntactic description* [Vol. 1]. 2nd ed. Cambridge: Cambridge University Press. pp.362–446.

———. 2008. “The place of Philippine languages in a typology of voice systems”. In Peter K. Austin and Simon Musgrave (eds.) *Voice and grammatical relations in Austronesian languages*. Stanford: CSLI Publications. pp.22–44.

Givón, Talmy. 1976. “Topic, pronoun, and grammatical agreement”. In Charles N. Li (ed.) *Subject and topic*. New York: Academic Press. pp.149–188.

———. 1979. “From discourse to syntax: Grammar as a processing strategy”. In Talmy Givón (ed.) *Discourse and syntax*. New York: Academic Press. pp.81–112.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Gundel, Jeanette K. Nancy Hedberg and Ron Zacharski. 1993. “Cognitive status and the form of referring expressions in discourse”. *Language* 69(2). pp.274–307.

Haig, Geoffrey and Stefan Schnell. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators (version 7.0)*. (https://lac.uni-koeln.de/en/multicast/) (Accessed 2018-03-01.)

——— (eds.). 2016[2015]. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. (https://lac.uni-koeln.de/multicast/) (Accessed 2018-03-01.).

———. 2016. “The discourse basis of ergativity revisited”. *Language* 92(3). pp.591–618.

Hopper, Paul. 1988. “Emergent grammar”. In Michael Tomasello (ed.) *The new psychology of language: Cognitive and functional approaches to language structure*. Mahwah, NJ: Erlbaum. pp.155–175.

Kumagai, Yoshiharu. 2006. “Information management in intransitive subjects: Some implications for the preferred argument structure theory”. *Journal of Pragmatics* 38(5). pp.670–694.

Kurabe, Keita. 2018. “The GRAID-annotated Jinghpaw corpus: Annotations and initial findings”. *Asian and African Languages and Linguistics* 12. pp.37–73.

Lyons, John. 1968. *Introduction to theoretical linguistics*. Cambridge: Cambridge University Press.

Mosel, Ulrike and Stefan Schnell. 2015. “Multi-CAST Teop”. In Geoffrey Haig and Stefan Schnell (eds.) *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. (https://lac.uni-koeln.de/multicast/) (Accessed 2018-03-01.)

Neeleman, Ad and Kriszta Szendröi. 2007. "Radical pro drop and the morphology of pronouns". *Linguistic Inquiry* 38(4). pp.671–714.

Noonan, Michael. 2003. *A crosslinguistic investigation of referential density.*. Unpublished manuscript. (http://crossasia-repository.ub.uni-heidelberg.de/190/) (Accessed 2018-03-01.)

Prince, Ellen F. 1981. "Toward a taxonomy of given-new information". In Peter Cole (ed.) *Radical pragmatics*. New York: Academic Press. pp.223–255.

Riesberg, Sonja and Beatrice Primus. 2015. "Agent prominence in symmetrical voice languages". *Language Typology and Universals (STUF)* 68(4). pp.551–564.

Riester, Arndt and Stefan Baumann. 2017. *The RefLex scheme — Annotation guidelines*. (SinSpeC: Working papers of the SFB 732 14.) Stuttgart: University of Stuttgart. (http://elib.uni-stuttgart.de/handle/11682/9028) (Accessed 2018-03-01.)

Schiborr, Nils N. 2015. "Multi-CAST English". In Geoffrey Haig and Stefan Schnell (eds.) *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. (https://lac.uni-koeln.de/multicast/) (Accessed 2018-03-01.)

Schiborr, Nils N., Stefan Schnell and Hanna Thiele. 2018. *RefIND — Referent Indexing in Natural-language Discourse: Annotation guidelines (v1.0)*. Unpublished manuscript.

Schnell, Stefan. 2015. "Multi-CAST Vera'a". In Geoffrey Haig and Stefan Schnell (eds.) *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. (https://lac.uni-koeln.de/multicast/) (Accessed 2018-03-01.)

Stoll, Sabine and Balthasar Bickel. 2009. "How deep are differences in referential density?". In Jiansheng Guo, Elena Lieven, Nancy Budwig, and Susan Ervin-Tripp, Keiko Nakamura and Seyda Özçaliskan (eds.) *Crosslinguistic approaches to the psychology of language: Research in the tradition of Dan Isaac Slobin*. London: Psychology Press. pp.543–555.

Van Valin, Robert D., Jr. and Randy J. LaPolla,. 1997. *Syntax: Structure, meaning, and function*. Cambridge: Cambridge University Press.

Appendix

List of GRAID symbols

Form symbols

⟨0⟩	contrastively suppressed argument position ("zero")
⟨f0⟩	[optional] implied participant in a non-finite clause
⟨pro⟩	free pronoun in its full form, in contrast to affixed ⟨-pro/pro-⟩ and cliticized ⟨=pro/pro=⟩ expressions
⟨np⟩	lexical NP
⟨refl⟩	overt reflexive or reciprocal pronoun
⟨w⟩	weak form, indicates phonologically lighter form of a particular element, that may under certain conditions be realized as a clitic; attaches to other form glosses, e.g. ⟨wpro⟩
⟨v⟩	lexical verb as the form element of a predicate
⟨vother⟩	verbal element, may be used in predicative function, but lacks the normal means for assigning arguments (e.g. imperatives, certain types of nominalizations)
⟨ln⟩	NP-internal subconstituent occurring left of the NP head
⟨rn⟩	NP-internal subconstituent occurring right of the NP head
⟨lv⟩	subconstituent of a verb complex occurring left of the verbal head
⟨rv⟩	subconstituent of a verb complex occurring right of the verbal head
⟨other⟩	form not relevant

Semantic person-animacy symbols

⟨.1⟩	argument with 1st person referent(s)
⟨.2⟩	argument with 2nd person referent(s)
⟨.h⟩	argument with 3rd person human referent(s)
ø	null gloss: argument with 3rd person non-human referent(s)
⟨.d⟩	[optional] argument with 3rd person anthropomorphized referent(s)

Function symbols

⟨:s⟩	intransitive subject
⟨:a⟩	transitive subject
⟨:ncs⟩	non-canonical subject, an argument that lacks some or all of the morphological properties associated with subjects, but commands most of the syntactic properties associated with subjects in the language concerned
⟨:p⟩	transitive object
⟨:g⟩	goal argument of a goal-oriented verb of motion, transitive or intransitive; also: recipients and addressees
⟨:l⟩	locative argument of verbs of location
⟨:obl⟩	oblique argument, excluding those glosses ⟨:g⟩ or ⟨:l⟩
⟨:poss⟩	possessor
⟨:appos⟩	apposition
⟨:dt⟩	dislocated topic
⟨:voc⟩	vocative, used for expressions denoting the person to which an utterance is addressed
⟨:pred⟩	predicate of a clause
⟨:predex⟩	predicate of an existential expression
⟨:other⟩	function not relevant

Miscellaneous symbols

⟨aux⟩	auxiliary
⟨cop⟩	overt copular verb, usually in combination with a non-verbal predicate complement
⟨adp⟩	adposition
⟨nc⟩	not considered / non-classifiable

Clause boundary symbols

⟨##⟩	left-edge boundary of a syntactically independent clause
⟨#⟩	left-edge boundary of all other clauses
⟨ds⟩	direct speech; attaches to ⟨##⟩, ⟨#⟩
⟨rc⟩	relative clause; attaches to ⟨#⟩
⟨ac⟩	adverbial clause; attaches to ⟨#⟩
⟨cc⟩	complement clause; attaches to ⟨#⟩
⟨.neg⟩	negated clause
⟨%⟩	right-edge boundary of an embedded clause, omitted if followed by ⟨##⟩ or ⟨#⟩